Actuarial

Data Analytics / Predictive Modeling Seminar Part 1

Wifi details:

ILTCI Conference

ILTCI2016

ILTCI

Actuarial

Data Analytics / Predictive Modeling Seminar Part 1

Jim Berger
Matt Morton
Ben Williams
Paul Bailey

ILTCI

16th Annual Intercompany Long Term Care Insurance Conference

Disclaimer

The data used in this seminar is fictitious, and it or results of analyses based on it should not be relied upon for any purpose

Agenda of the Seminar

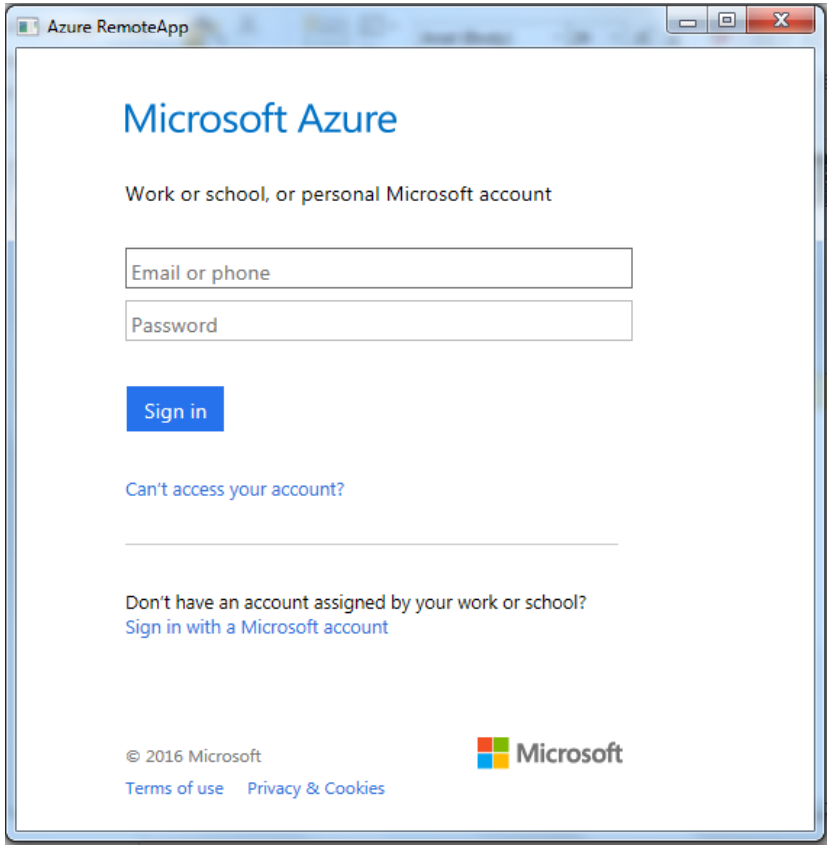
1. Background to the Workshop
2. Predictive Modeling and Applications to Long Term Care Insurance
3. Predictive Modeling of Long Term Care Incidence using Generalized Linear Models in Emblem

Agenda of today's session

- Towers Watson Emblem
  - What is it and why are we using it for the workshop?
  - Ensuring everyone has access to it
- What predictive modeling data looks like
- Preliminary Analyses in Emblem using LTC incidence data
  - Correlation analysis
  - Univariate/bivariate analysis
  - Distribution of response (discussion only)
- Statistical Background to Generalized Linear Models (GLMs)

Logging on to Microsoft Azure

Open The Azure Remote App:



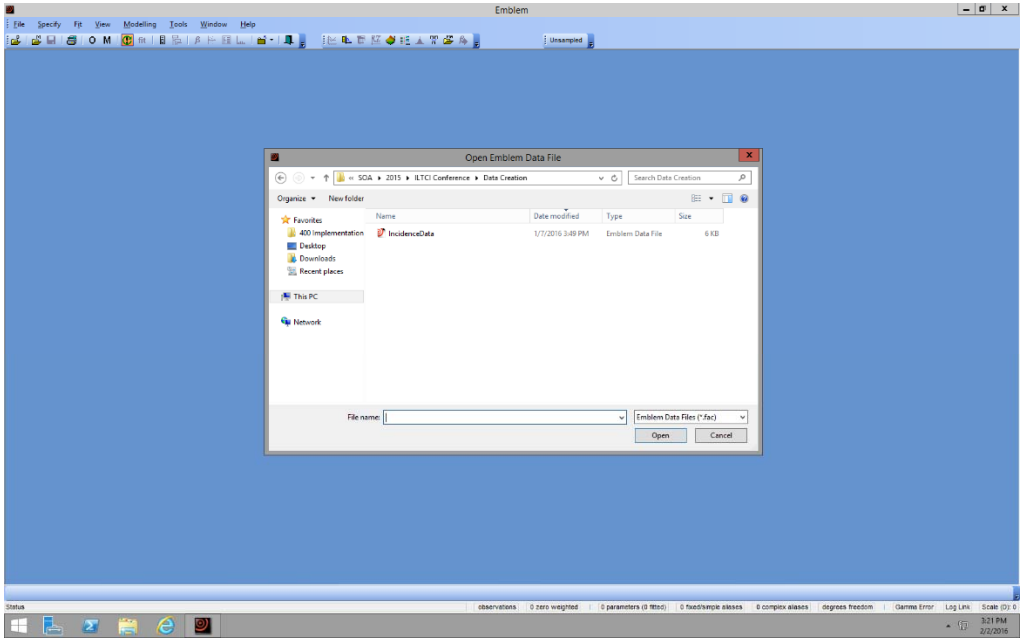
Login: TrainXX@wtwsaas.com

Password: T0wersW@tson16

where XX is the number written on your handout

Towers Watson Emblem

- What is it?
  - Market-leading GLM-fitting software
- Why are we using it in this seminar?
  - It is easy to use (no programming required), fast and visual
- Ensuring everyone has access to it:
  - Log on to Azure, taking note of your username and password
  - If you can open Emblem and get to a screen like the one shown, you are good to go



What Predictive Modeling Data looks like

What is Predictive Modeling?

- We have a series of historical observations showing how input variables x (independent variables) are related to responses y (dependent variables) by some process

- A predictive model is some function of the independent variables that mimics this process
- Predictive Modeling is the process of developing a predictive model
- There are two goals in fitting a predictive model
  - Prediction: to be able to predict responses for future input variables x
  - Information: to understand the process that associates response variables with input variables
- Note: Just the same as the goal of developing an assumption!!!

What Predictive Modeling Data looks like

- The data used reflects this
- Each historical observation is a row in the data
- The dependent response variable y and each independent variable x are columns in the data
- It is common to have a column of weights indicating how much attention should be paid to each observation when fitting the model
  - For example, a policy exposed for 1 year will have twice the influence of a policy exposed for six months, if exposure is chosen as the weight
- A sample taken from the incidence modeling data is included (IncidenceDataSample.xlsx)
  - For this data, (claims/years of exposure) is the response, years of exposure is the weight
- Creating model ready data can be non-trivial

Preliminary Analyses in Emblem using LTC incidence data

Data exploration is carried out to ensure that the data is appropriate for modeling, i.e.

- All relevant independent variables are present and correctly populated
- The weight is distributed appropriately among the levels of the independent variables
- The response varies appropriately among the levels of the independent variables
- The relationships between the independent variables are reasonable and understood
- The response is appropriately distributed

Preliminary Analyses in Emblem using LTC incidence data

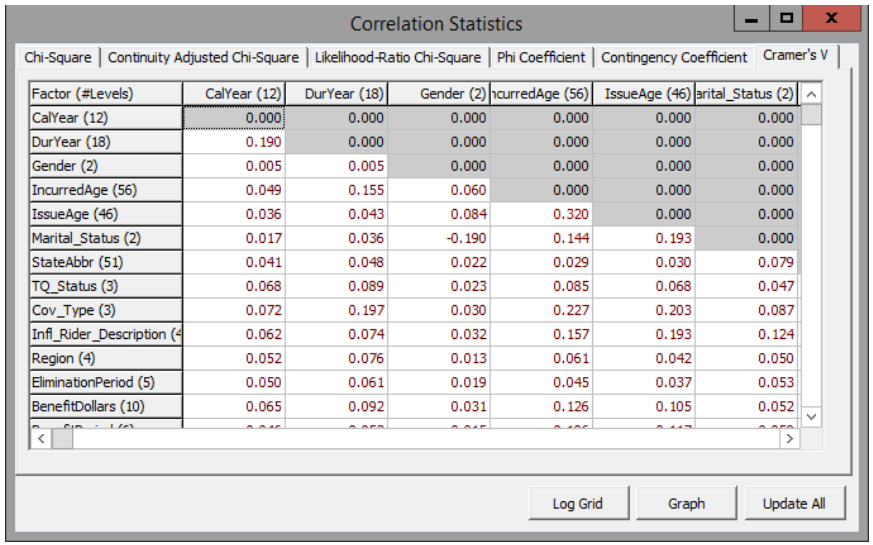
This is done by looking at the following

- Summaries of response and weight by combinations of independent variables
- Analysis of correlations between independent variables
- Analysis of distribution of the response variable

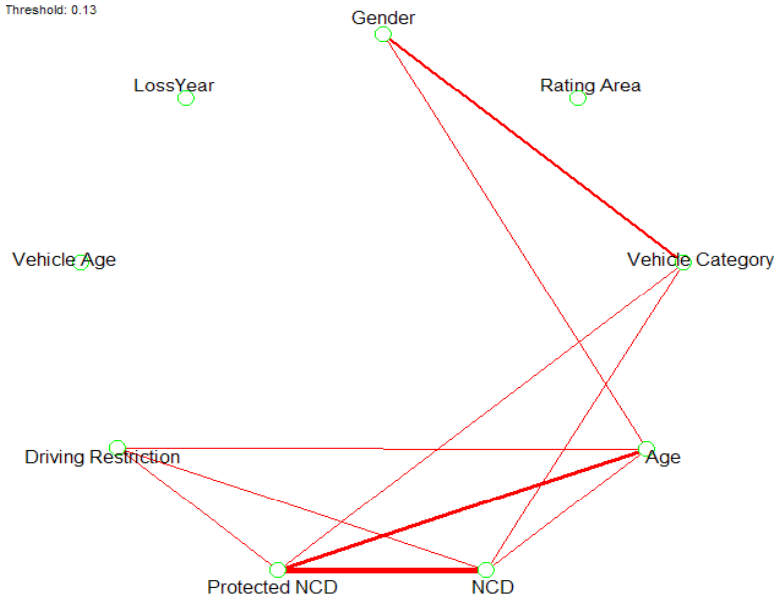
Preliminary Analyses in Emblem using LTC incidence data

Within Emblem,

- Open the Incidence data (File>Open Data File) and press Enter
- Run Correlations Analysis (Modelling>Initial Tests>Correlations)
- What do you notice? Does it make sense?



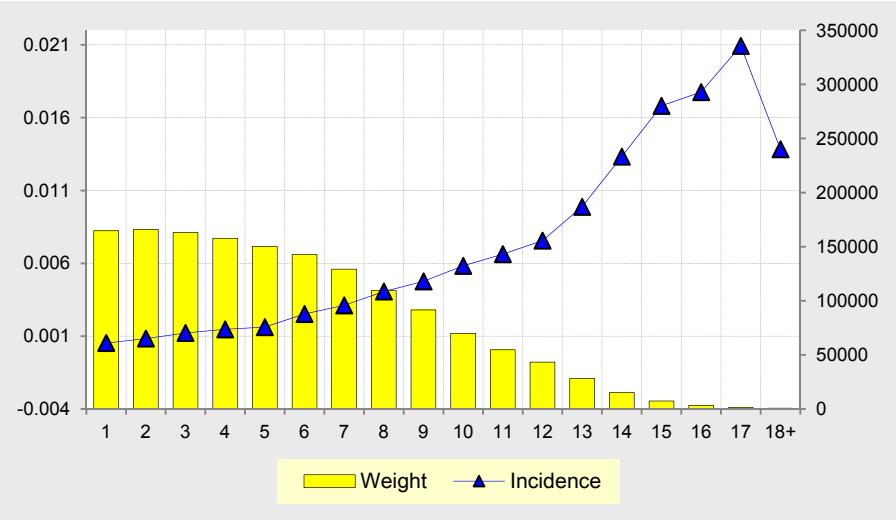
Factor (#Levels)	CalYear (12)	DurYear (18)	Gender (2)	IncurredAge (56)	IssueAge (46)	Marital_Status (2)
CalYear (12)	0.000	0.000	0.000	0.000	0.000	0.000
DurYear (18)	0.190	0.000	0.000	0.000	0.000	0.000
Gender (2)	0.005	0.005	0.000	0.000	0.000	0.000
IncurredAge (56)	0.049	0.155	0.060	0.000	0.000	0.000
IssueAge (46)	0.036	0.043	0.084	0.320	0.000	0.000
Marital_Status (2)	0.017	0.036	-0.190	0.144	0.193	0.000
StateAbbr (51)	0.041	0.048	0.022	0.029	0.030	0.079
TQ_Status (3)	0.068	0.089	0.023	0.085	0.068	0.047
Cov_Type (3)	0.072	0.197	0.030	0.227	0.203	0.087
Infl_Rider_Description (4	0.062	0.074	0.032	0.157	0.193	0.124
Region (4)	0.052	0.076	0.013	0.061	0.042	0.050
EliminationPeriod (5)	0.050	0.061	0.019	0.045	0.037	0.053
BenefitDollars (10)	0.065	0.092	0.031	0.126	0.105	0.052

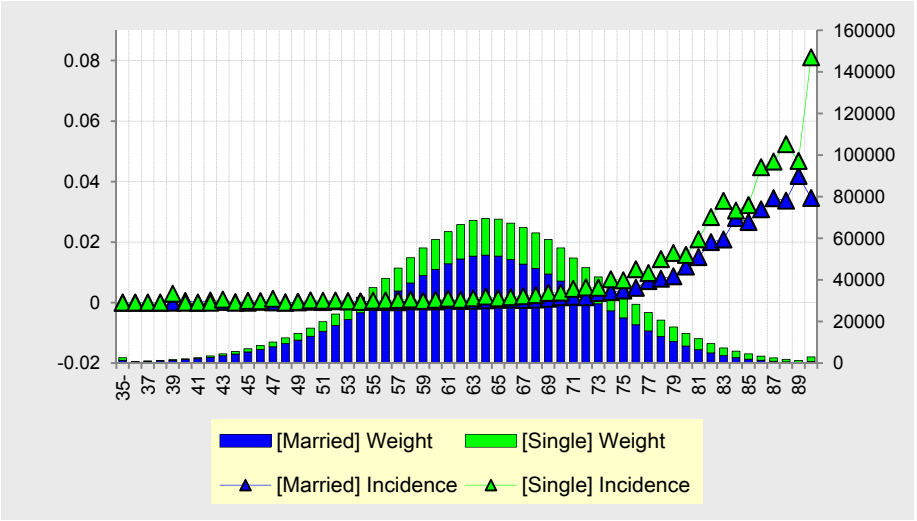


Preliminary Analyses in Emblem using LTC incidence data

Within Emblem,

- Open Data Analyzer (Modeling>Initial Tests>Data Analyzer)
- Create some Univariate and Bivariate Analyses based on different independent variables
- Do you note anything that surprises you?



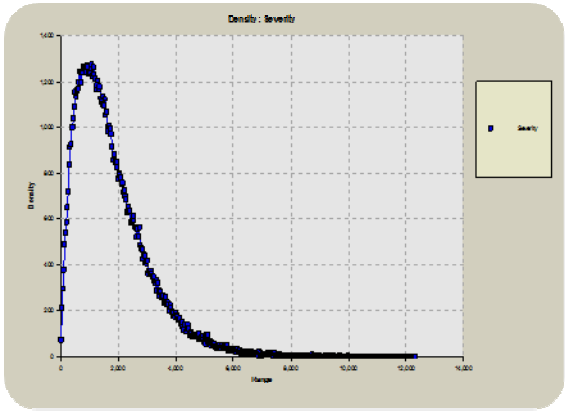


Preliminary Analyses in Emblem using LTC incidence data

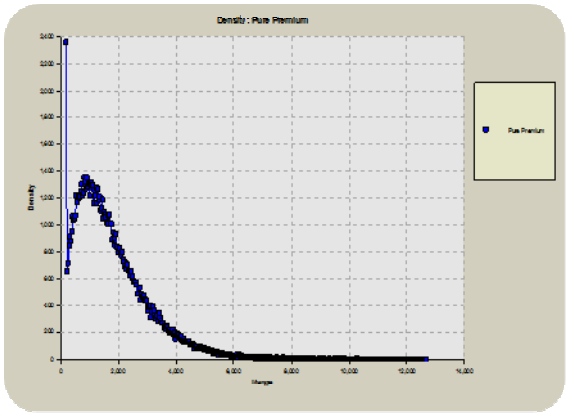
In some cases it is important to review the distribution of the response, in order to

- Make sure that it is appropriate, for example
  - Claim costs should be positive, with an appropriate mean, median etc.
  - Proportion of benefits should be between 0 and 1
- Help guide our selection of error structure
  - If response is between 0 and 1, Binomial is best choice
  - If response is negative, Gamma is a bad choice
- This is not useful for the response in the Incidence data

Consistent with Gamma

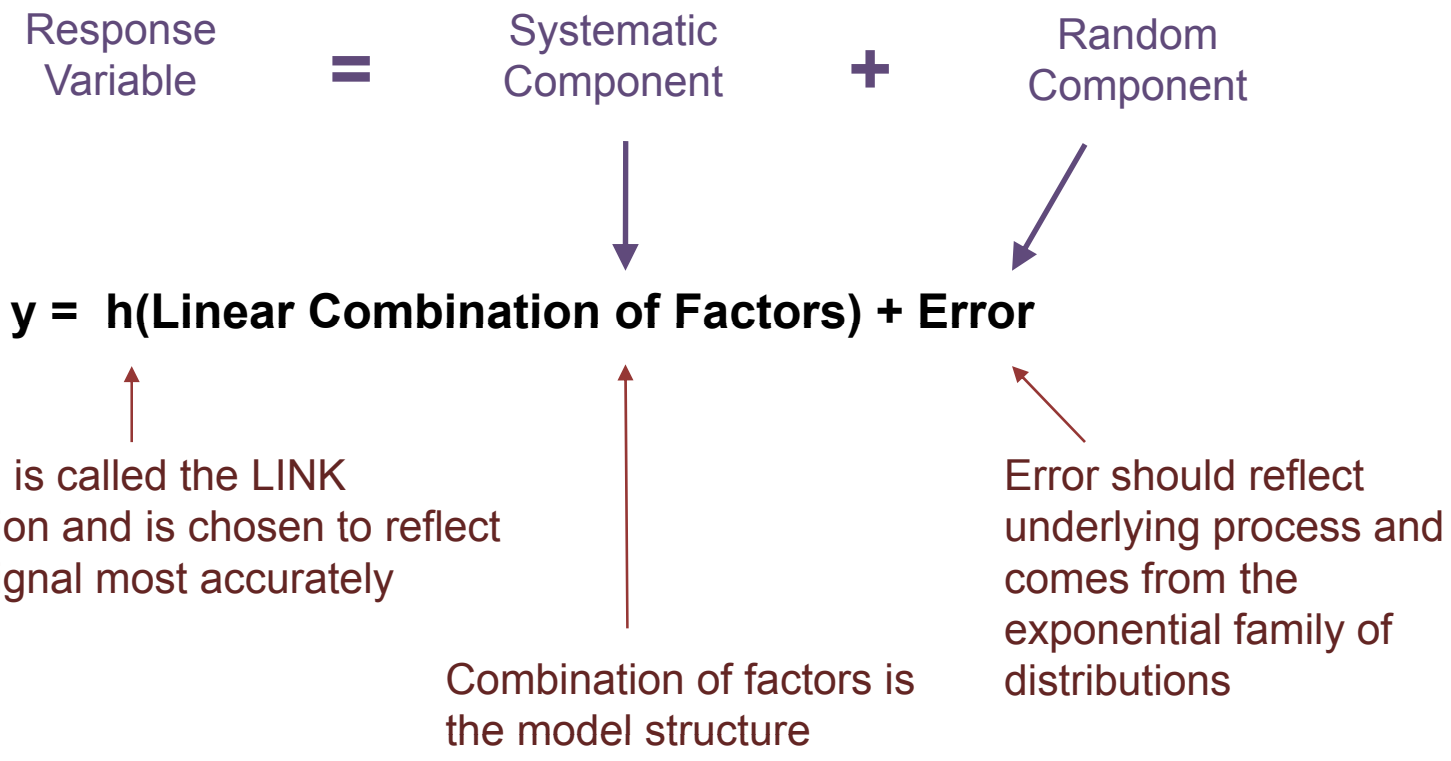


Consistent with Tweedie



Statistical Background to GLMs

GLMs (Generalized Linear Models) are a flexible, transparent and sophisticated predictive modeling technique



GLM Building Blocks

y = **h(**Combination of Factors**)** + Error

- **Link function** ($g=h^{-1}$) choice determines how the factors are related to the response in the model
  - Identity: Variables related additively (e.g., risk modeling)
  - Log: Variables related multiplicatively (e.g., risk modeling)
  - Logit: Modeling binary outcomes (e.g. yes/no events)

- Note that if link function is log, then
  - Inverse of Log is Exponential
  - Exp(A+B+…G) = Exp(A).Exp(B)..Exp(G)

  Model is multiplicative

GLM Building Blocks

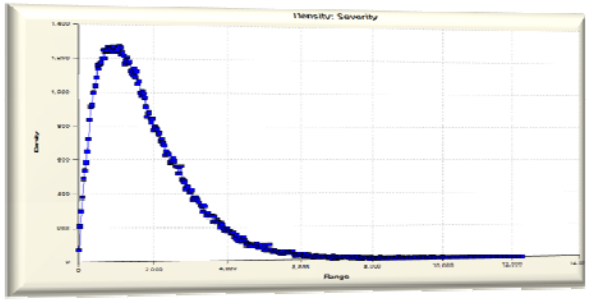
y = h(**Combination of Factors**) + Error

- Include variables that are predictive; exclude those that are not
  - Gender may not have major impact on utilization
- Simplify some variables, if full inclusion is not necessary
  - Some levels within a particular predictor may be grouped together e.g. those with incurred age under 50, where incidence is very practically zero
  - A curve may replicate the effect of an ordinal variable e.g. incurred age or duration
- Complicate model structure if the responses depends on combinations of levels of more than one variable (interactions)
  - E.g. the difference between males and females depends on incurred age

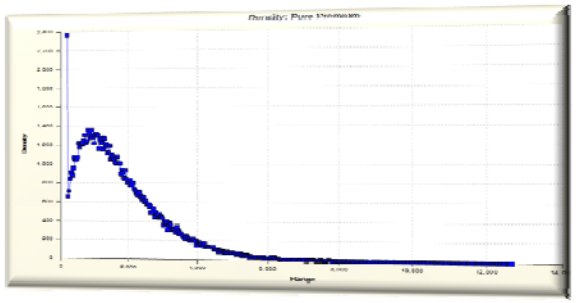
GLM Building Blocks

y = h(Combination of Factors) + Error

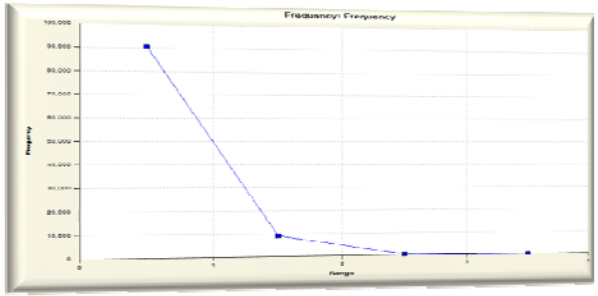
- Error functions reflects the variability of the underlying process and can be any distribution within the exponential family of distributions, for example:



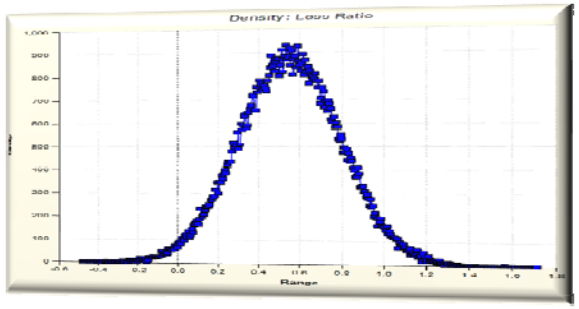
- Gamma consistent with severity modeling



- Tweedie consistent with pure premium modeling



- Poisson consistent with incidence modeling



- Normal useful for a variety of applications

- Appropriate for modeling processes common in insurance

Statistical Background to GLMs

- Generally accepted standards are good starting points for link functions and error structures

Observed Response	Most Appropriate Link Function	Most Appropriate Error Structure
--	--	Normal
Frequency/Mortality/Incidence	Log	Poisson
Severity/Utilization	Log	Gamma
Pure Premium	Log	Tweedie
Retention/Conversion/Termination/ Cross-sell/Response Rate	Logit	Binomial

What we have covered

- Towers Watson Emblem
  - What is it and why are we using it for the workshop?
  - Ensuring everyone has access to it
- What predictive modeling data looks like
- Preliminary Analyses in Emblem using LTC incidence data
  - Correlation analysis
  - Univariate analysis
  - Bivariate analysis
  - Distribution of response (discussion only)
- Statistical Background to Generalized Linear Models (GLMs)

Reminder of further sessions

1. Background to the Workshop
2. **Predictive Modeling and Applications to Long Term Care Insurance**
3. **Predictive Modeling of Long Term Care Incidence using Generalized Linear Models in Emblem**

We'll see you on Wednesday, when we will use Emblem to fit a GLM of the incidence data explored today!